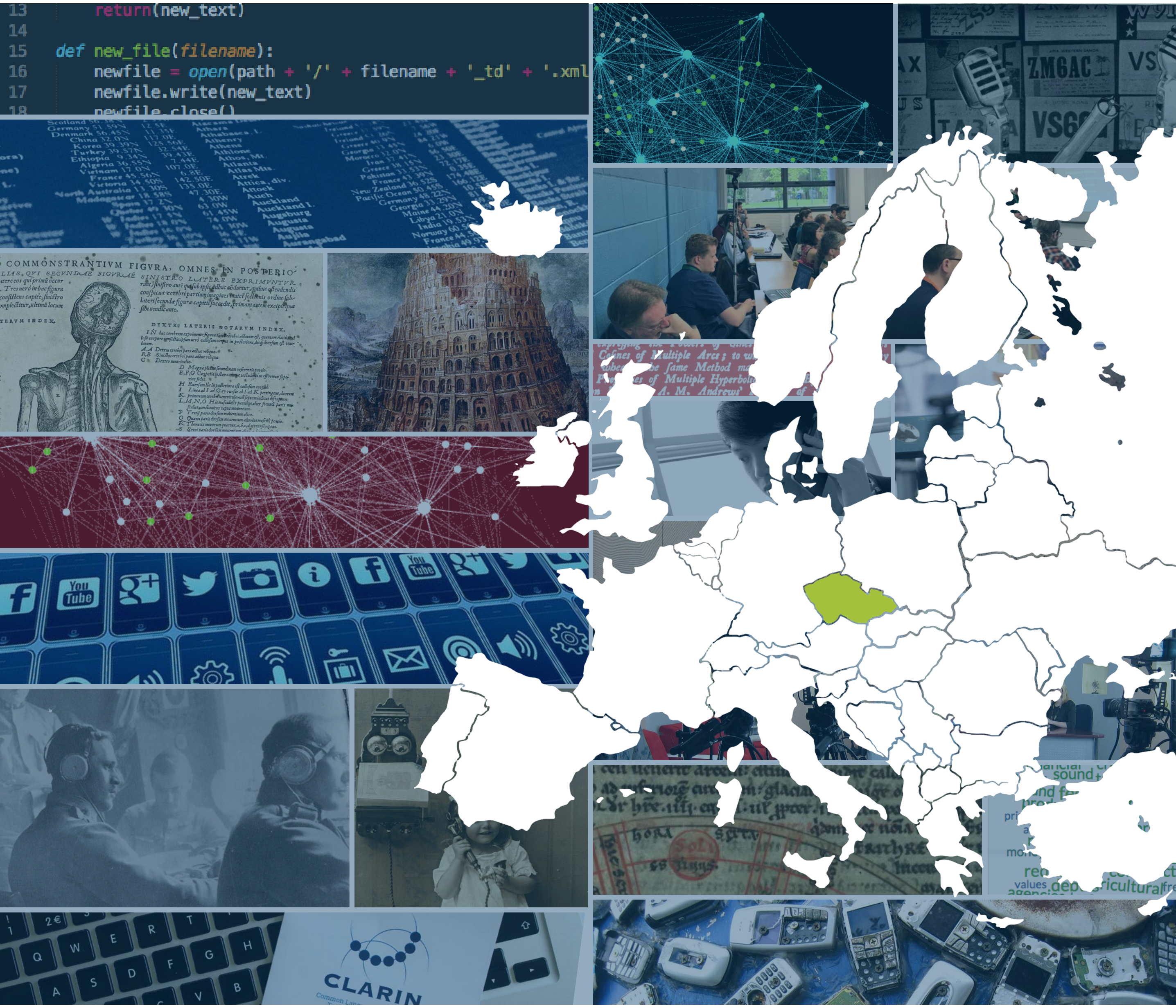
CLARIN

Common Language Resources and Technology Infrastructure

Tour de CLARIN

Czech Republic



Written by Barbora Hladka, Darja Fišer and Jakob Lenardič, and edited by Darja Fišer and Jakob Lenardič

Foreword

Tour de CLARIN highlights prominent user involvement activities of CLARIN national consortia with the aim to increase the visibility of CLARIN consortia, reveal the richness of the CLARIN landscape, and display the full range of activities throughout the CLARIN network that can inform and inspire other consortia as well as show what CLARIN has to offer to researchers, teachers, students, professionals and the general public interested in using and processing language data in various forms.

This brochure presents Czech Republic and is organized in five sections:

- Section One presents the members of the consortium and their work
- Section Two demonstrates an outstanding tool
- Section Three highlights a prominent resource
- Section Four reports on a successful event for researchers and students
- Section Five includes an interview with a renowned researcher from the digital humanities or social sciences who has successfully used the consortium’s infrastructure in their research

Prague, Czech Republic | photo by Rodrigo Ardilha | Unsplash

Czech Republic

Written by Darja Fišer and Jakob Lenardič

The Czech consortium LINDAT[1] is a founding member of CLARIN ERIC. It is a B-certified centre that involves four Czech research institutions – the Department of Cybernetics at the University of West Bohemia, the Institute of Formal and Applied Linguistics at Charles University, the Czech Language Institute at the Czech Academy of Science, and the NLP Centre at Masaryk University. The consortium is led by Professor Eva Hajičová.

The consortium offers a pioneering repository for language resources, whose architecture serves as the backbone of several other CLARIN repositories. The repository rigorously follows best practices on metadata presentation, so it is ensured that all language data are safely stored with clear documentation as well as outfitted with guidelines on proper citation. Many of the monolingual, parallel and speech corpora within the repository can be accessed through the concordancer KonText, which is a flexible search environment that allows users to perform queries of various complexities – from simple searches by lemma or word form to using CQL – as well as save search results for future research.

LINDAT also offers an integrated environment for storing, building, searching and visualising treebanks, which are databases of syntactically annotated sentences. As a pivotal tool for treebanks, LINDAT offers PLM Tree Query, through which researchers can browse a great variety of treebanks in 61 languages. For the novice researcher, the Tree Query is accompanied by a step-by-step tutorial that shows how to execute searches in the query language. Together with the Norwegian INESS, LINDAT is a CLARIN Knowledge Centre that specialises in the creation and maintenance of treebanks.

LINDAT Team | Back row: Pavel Straňák, Jaroslava Hlaváčová, Pavel Pecina, David Mareček, Ondřej Bojar, Jan Hajič, Milan Fučík. Front row: Barbora Hladká, Anna Nedoluzhko, Vendula Kettnerová, Eva Hajičová (national coordinator), Anna Vernerová, Veronika Kolářová, Magda Ševčíková, Marie Křížková.

LINDAT actively works on introducing its state-of-the-art language technologies to researchers both within computational fields like NLP and within the digital humanities and social sciences. To this end, LINDAT organised a user involvement workshop on 24 April 2018 in Prague, which aimed to showcase how technological infrastructures are also relevant beyond the computational framework. You can read more about the workshop on page 10.

Prague, Czech Republic | photo by Studio Reasons | Unsplash

[1] https://lindat.mff.cuni.cz/en

UDPipe

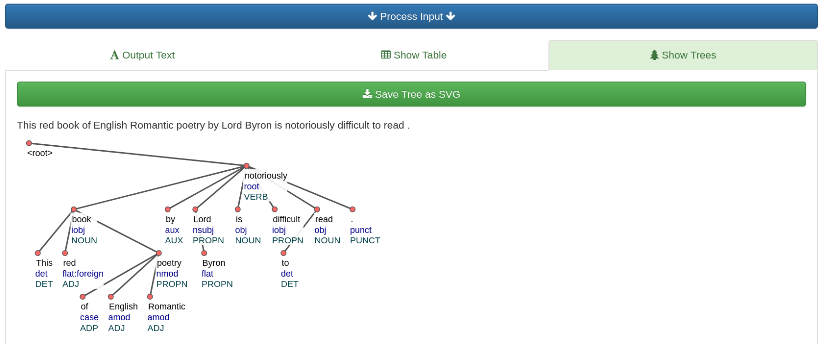
Written by Barbora Hladka and Jakob Lenardič, edited by Darja Fišer

UDPipe[2] is a state-of-the-art tool pipeline which performs several complex annotation tasks: tokenisation, Part-of-Speech tagging, lemmatisation, sentence segmentation and dependency parsing, all to a high degree of precision. The architecture of UDPipe employs a deep neural network and is trained on language models from the Universal Dependency treebanks provided by LINDAT (see page 8 for a presentation of the Universal Dependencies). UDPipe can be used to annotate and parse texts from over 50 languages, many of which are non-Indo-European, such as Arabic, Irish, Indonesian and Tamil. It was (and is being) developed at the Institute of Formal and Applied Linguistics at Charles University, and can be freely used for non-commercial purposes.

UDPipe is available both as a downloadable program that is compatible with Linux, Windows and OS X, as a library in programming languages such as C++, Python, Perl, R, Java, C#, and as an easy-to-use web application. Researchers who wish to run UDPipe as a standalone program on their own computers must also download one of the Universal Dependencies language models, which are described in detail in the UDPipe User's Manual:

- the Universal Dependencies 1.2 models, which contain cross-linguistically consistent treebank annotation models for 33 languages;
- the Universal Dependencies 2.0 models, which are an updated version of the former and contain annotation models for over 50 languages; and
- the CoNLL17 Shared Task Baseline UD 2.0 models, which contain a different version of the Universal Dependencies 2.0 models.

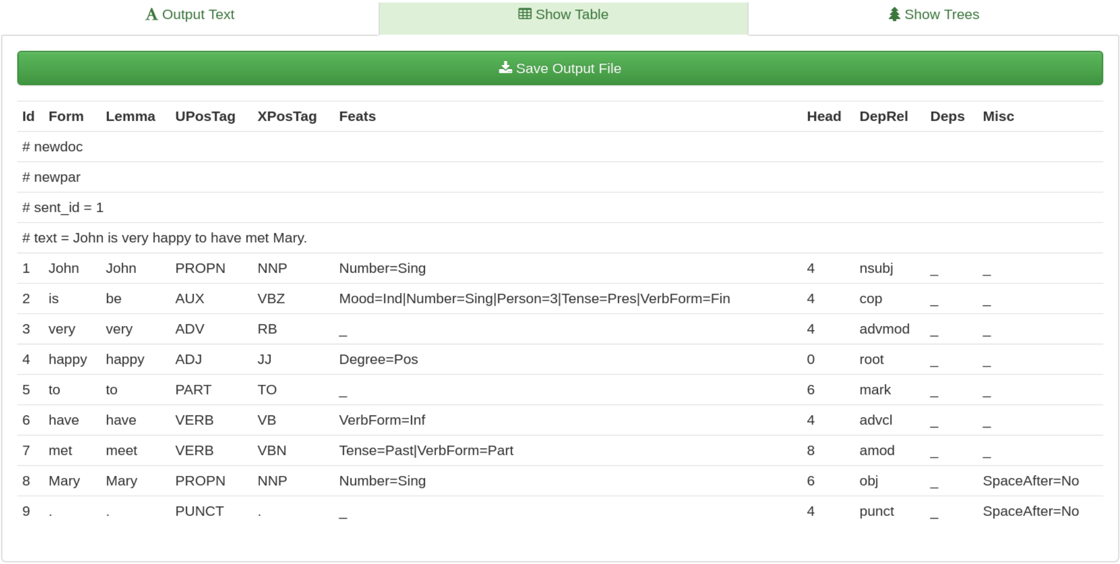
The UDPipe Web Application is provided through the LINDAT architecture. It is very easy to use in the sense that researchers need only select one of the many languages in one of the three training models and input the text (or upload whole files) they wish to have annotated. The results can either be visualised in the form of a tree structure, which shows the syntactic dependencies (Figure 1), or in table form, where each individual word is accompanied by its Part-of-Speech label as well as more complex set of grammatical features, such as case, person, gender, and tense (Figure 2).



This red book of English Romantic poetry by Lord Byron is notoriously difficult to read .

Figure 1: The tree structure of a complex English raising construction. Apart from visualising the sentential structure, the tree structure also shows the parts of speech and syntactic features of the constituents.

[2] http://lindat.mff.cuni.cz/services/udpipe/



Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# newdoc									
# newpar									
# sent_id = 1									
# text = John is very happy to have met Mary.									
1	John	John	PROPN	NNP	Number=Sing	4	nsubj	_	_
2	is	be	AUX	VBZ	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	4	cop	_	_
3	very	very	ADV	RB	_	4	advmod	_	_
4	happy	happy	ADJ	JJ	Degree=Pos	0	root	_	_
5	to	to	PART	TO	_	6	mark	_	_
6	have	have	VERB	VB	VerbForm=Inf	4	advcl	_	_
7	met	meet	VERB	VBN	Tense=Past|VerbForm=Part	8	amod	_	_
8	Mary	Mary	PROPN	NNP	Number=Sing	6	obj	_	SpaceAfter=No
9	.	.	PUNCT	.	_	4	punct	_	SpaceAfter=No

Figure 2: UDPipe shows the grammatical features of the sentence "John is very happy to have met Mary" in table form. Note that it can detect very complex features, such as the perfect (i.e. past tense) use of the infinitive in the subordinate clause.

The powerful flexibility of UDPipe was demonstrated in the CoNLL 2017 shared task, which was of crucial importance for the development and research of dependency parsing. In the shared tasks, UDPipe was used to process raw text in 40+ languages based on the Universal Dependency models with very high precision, which shows that UDPipe can also be easily adapted to annotate and parse new languages. The CoNLL 2018 is a follow-up of CoNLL 2017 and UDPipe is used as a baseline system.

For more details on UDPipe see Straka and Straková (2017) and Straka et al. (2016):

Straka, M. and Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with Udpipe. In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017. http://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf

Straka, M., Hajič, J., and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 2016. http://ufal.mff.cuni.cz/~straka/papers/2016-lrec_udpipe.pdf

Universal Dependencies (UD)

Written by Barbora Hladka, edited by Darja Fišer and Jakob Lenardič

Universal Dependencies (UD)[3] is an open collaboration project in the field of Natural Language Processing (NLP). Its motivation comes from multi- and cross-lingual research, and its goal is to develop a universal approach to grammatical annotation, applicable to as many languages as possible. UD is administered by an international team under supervision of Joakim Nivre. The UD project has been up and running since the spring of 2014.

UD provides a universal inventory of part-of-speech categories and syntactic relations for consistent cross-linguistic annotation, as well as several existing treebanks that are richly annotated with the grammatical features. The following picture shows a UD tree structure for the sentence "Mary loves John". Three part-of-speech categories – PROPN (proper name), VERB (verb), and PUNCT (punctuation) – and four syntactic relations – root (predicate), nsubj (nominal subject), obj (object), and punct (punctuation) – occur in the tree.

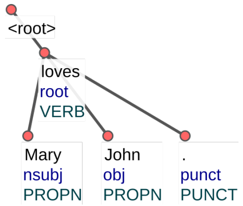
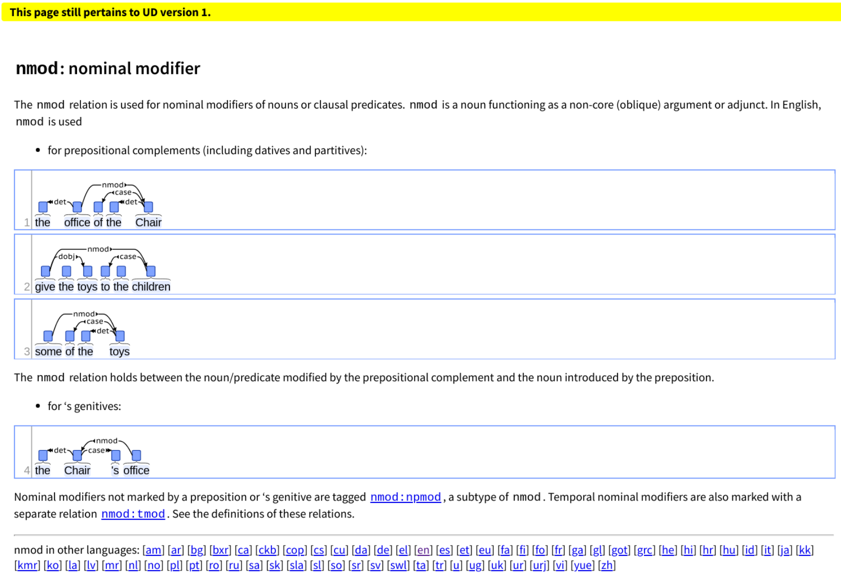


Figure 3: A syntactic tree for "Mary loves John".

UD is also accompanied by detailed guidelines for carrying out the annotation, with examples from numerous languages. The following figure illustrates the complex criteria UD uses to recognise nominal modifiers, which often also take into account complex grammatical interdependencies from formal grammar, such as case assignment/checking.



This page still pertains to UD version 1.

nmod: nominal modifier

The nmod relation is used for nominal modifiers of nouns or clausal predicates. nmod is a noun functioning as a non-core (oblique) argument or adjunct. In English, nmod is used

- for prepositional complements (including datives and partitives):

1 the office of the Chair

2 give the toys to the children

3 some of the toys

The nmod relation holds between the noun/predicate modified by the prepositional complement and the noun introduced by the preposition.

- for 's genitives:

4 the Chair 's office

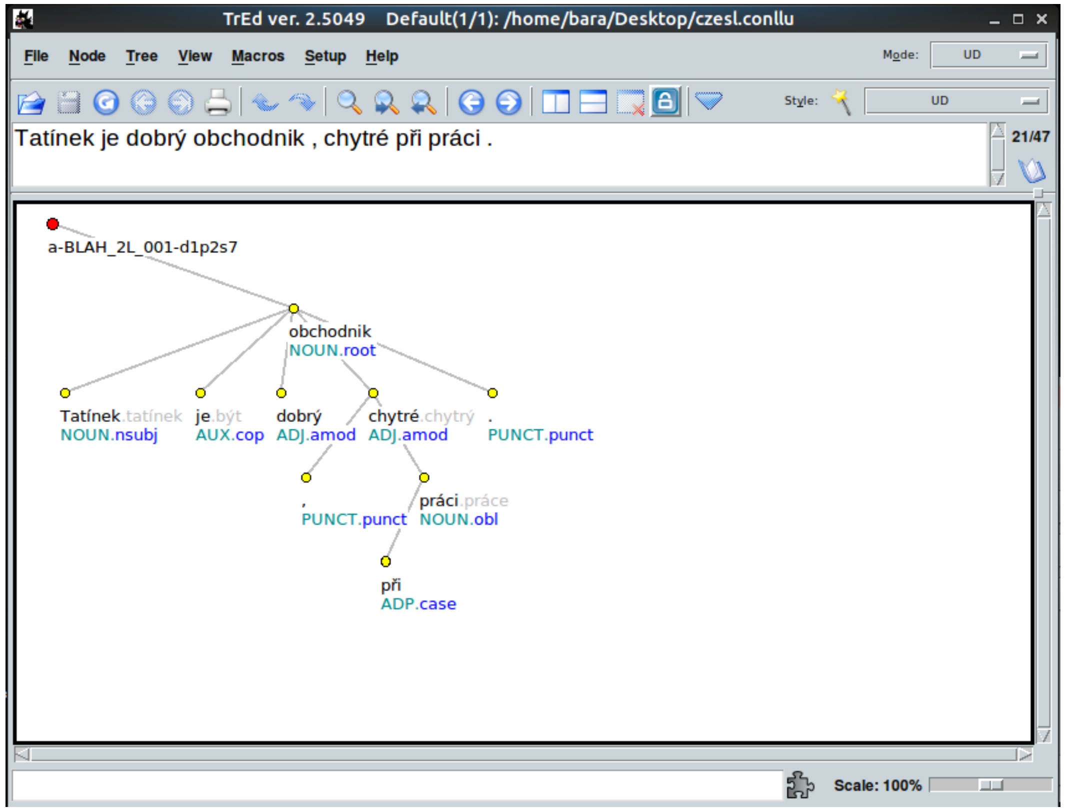
Nominal modifiers not marked by a preposition or 's genitive are tagged nmod:npmod, a subtype of nmod. Temporal nominal modifiers are also marked with a separate relation nmod:tmod. See the definitions of these relations.

nmod in other languages: [am] [ar] [bg] [bxr] [ca] [ckb] [cop] [cs] [cu] [da] [de] [el] [en] [es] [et] [eu] [fa] [fi] [fo] [fr] [ga] [gl] [got] [grc] [he] [hi] [hr] [hu] [id] [it] [ja] [kk] [kmr] [ko] [la] [lv] [mr] [nl] [no] [pl] [pt] [ro] [ru] [sa] [sk] [sla] [sl] [so] [sr] [sv] [swl] [ta] [tr] [u] [ug] [uk] [ur] [urj] [vi] [yue] [zh]

Figure 4: The definition of a nominal modifier in UD.

[3] http://universaldependencies.org/

To search the UD treebanks, researchers can use the online PML-TQ (PML Tree Query) service and UDPipe (presented on page 6), which is an automatic UD annotation pipeline that uses models trained for nearly all the treebanks, so it offers an easy access point to the Universal Dependencies. A number of graphical user interfaces for manual UD annotation are also available. One of them is TrEd, which is a fully customisable and programmable editor and viewer of tree structures developed at the Institute of Formal and Applied Linguistics. The editor, which offers an extension for UD annotation illustrated in the following picture, has been successfully used to annotate thousands of sentences in the Prague Dependency Treebanks.



TrEd ver. 2.5049 Default(1/1): /home/bara/Desktop/czesl.conllu

Tatínek je dobrý obchodník , chytré při práci .

Figure 5: Using the editor TrEd to parse a Czech sentence.

A new version of the UD treebanks is released every six months. The latest version (2.1) came out at the end of 2017 and consists of an impressive number of treebanks, 102, for an equally impressive number of languages, 60. This version offers a ten times greater number of treebanks for six times more languages than the very first release in 2014, which shows how the inclusion of new language data is growing exponentially. All the versions are downloadable from the LINDAT/CLARIN repository.

After a period of rapid growth in 2014–2017, LINDAT organised a series of events dedicated to training and conducting parsing experiments with UD treebanks, as well as discussions of UD-related topics. Among these was a tutorial on UD at the EACL 2017 conference in Valencia in Spain, the first workshop on UD in Gothenburg in Sweden in May 2017, and the CoNLL 2017 and 2018 Shared Tasks, in which the UD treebanks were successfully used as models for the development of advanced dependency parsers.

DARIAH-CZ Workshop on Digital Humanities 2018

Written by Barbora Hladka, edited by Darja Fišer and Jakob Lenardič

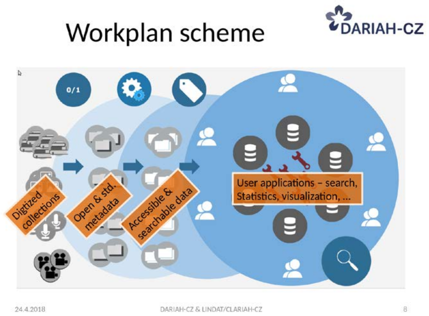
The members and partners of the Czech CLARIN consortium recently submitted a proposal to establish DARIAH-CZ, a Czech node of the DARIAH European researcher infrastructure for arts and humanities. In light of the proposal, a one-day international workshop titled DARIAH-CZ Workshop on Digital Humanities 2018[4] was held in Prague on 24 April 2018 at the Academy of Sciences of the Czech Republic in order to introduce the project and generally promote computational approaches within humanities and social sciences, both in the Czech Republic and internationally. During the workshop, sixteen lectures were given by prominent computational and digital humanities researchers working at leading Czech and European research institutions. The workshop was well attended. There were around 70 participants, most of whom were researchers from various Czech institutions while some also came from Slovakia, Poland, Hungary, and Germany.

The workshop began with the introduction of the European projects and, institutes related to the DARIAH-CZ project both structurally (DARIAH, DARIAH-PL, DESIR) and thematically (EADH, Austrian Centre for Digital Humanities). In the afternoon, the Czech projects that are planned to be integrated in DARIAH-CZ were presented. Perhaps most prominently, Pavel Straňák gave a comprehensive presentation of the LINDAT/CLARIN repository and Silvie Cinková introduced the recently established Czech Association for Digital Humanities, which will be a partner in the project. The afternoon session was concluded with a lecture on EHRI, which is a portal dedicated to the presentation and interpretation of Holocaust-related archival documents on the basis of digital tools.

In conclusion of the workshop, the following projects were presented to showcase the successful application of computational methods within the humanities and social sciences:

- the GEHIR project, which is an interdisciplinary research initiative that applies computational methods to the historiography of ancient Graeco-Roman religions;
- the Archaeological Information System of the Czech Republic, which is a tool used to integrate digital resources on Czech archaeology;
- the READ project and its main system, Transkribus, for transcribing and searching historical text collections; and
- Electronic Enlightenment, which is a wide-ranging online collection of edited correspondence from the early 17th to the mid-19th centuries.

The workshop successfully raised awareness of the proposed DARIAH-CZ and its related projects in the context of digital humanities. In addition, it strengthened the ties among the members of the Czech CLARIN consortium, its related partners and other national and international institutions, opening new research avenues for further collaboration.



Workplan scheme

DARIAH-CZ

Digtized collections

Open & std. metadata

Accessible & searchable data

User applications – search, Statistics, visualization, ...

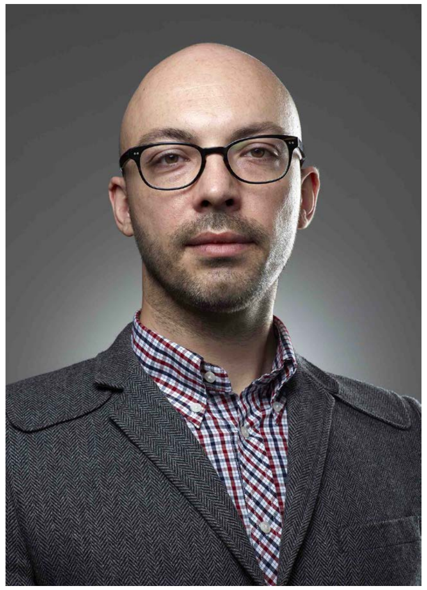
24.4.2018 DARIAH-CZ & LINDAT/CLARIAH-CZ 8

Figure 6: Boosting research in Digital Humanities with Research Infrastructures.

[4] https://www.lib.cas.cz/en/dariah-cz-workshop-2018/

Radim Hladík

Radim Hladík is a postdoctoral researcher at the Institute of Philosophy at Academy of Sciences of the Czech Republic in Prague and at the National Institute of Informatics in Japan. The following interview took place via Skype on 16 May 2018 and was conducted and transcribed by Jakob Lenardič, edited by Darja Fišer.



1. Please describe your academic background and your current position(s).

I received my PhD in sociology at the Faculty of Social Sciences in Prague. Currently, I'm a JSPS post-doctoral fellow at the National Institute of Informatics in Japan, where I am representing my sending organisation, the Institute of Philosophy of the Czech Academy of Sciences. Many of my colleagues in Japan are computer scientists, so this is a wonderful opportunity for me to improve my coding skills and be inspired about how to combine computational methodologies with social science research topics.

2. How did you get involved with Czech CLARIN consortium? Could you describe your collaboration with the consortium?

Two years ago, as a delegate of the Institute of Philosophy, I was one of the coordinators in a fairly large digital humanities project by the Library of the Czech Academy of Sciences. The proposal was ambitious, since we wanted to strike up a collaboration between many Czech institutes relevant for digital humanities, such as libraries, universities, and various institutes for linguistics and social sciences. Sadly, the project never left the planning stages, but it nevertheless brought together proponents of digital humanities, including me and the colleagues from Czech CLARIN. I was very inspired by their work and soon started learning how to code and apply computational approaches to my own research, which is otherwise rooted in sociology and media studies. Since then, I've been using tools and resources that Czech CLARIN provides and am in contact with their experts like Pavel Straňák, with whom I discuss my work and who has often helped me with technical issues.

As for concrete collaborations, we've recently established the Czech Association of Digital Humanities, for which I currently serve as the Chair. Several people from Czech CLARIN are very active in this association, like Eva Hajičová and Silvie Cinková. We've also submitted a project under the Czech DARIAH node last year with Czech CLARIN as the principal investigator. Its goal is to conduct an extensive corpus-based analysis of modern Czech texts from various domains (e.g., 20th century philosophy). I'll be involved as a representative of the Institute of Philosophy, which aims to contribute its historical and philosophical corpora and texts collections to Czech CLARIN. I believe that such a collaboration is of great importance for both sides. On the one hand, Czech CLARIN will give us an invaluable platform for the curation and sustainability of our resources, while on the other they'll be able to expand the applicability of their tools to new domains and across historical language variations based on our resources.

3. Which are the tools and resources provided by Czech CLARIN that you use in your research? Could you discuss how you use them in your own work?

If you work with texts in a language that is as morphologically complex as Czech, lemmatisation and morphosyntactic annotation of texts is needed even for the simplest analyses. In this sense, the tools that Czech CLARIN provides are essential for my current work.

I'd like to point out MorphoDiTa,[5] which is a tool for tokenisation, lemmatisation and morphological analysis. What I especially appreciate about MorphoDiTa is its flexibility, in that you don't need to install it as a stand-alone program on your computer, but you can use it as an API service which you easily integrate in your own code. This way, I don't need to worry about having additional components installed and their dependencies. I often come across tools that require a complicated installation processes, which dissuades me from using them.

What I also appreciate is that the Czech CLARIN repository keeps track of all the versions of a resource you upload. I believe this takes a lot of pressure off the whole publishing process, since I know that I can always publish a newer version of a specific dataset in case I do some additional work on it, making me more confident in releasing a dataset sooner, since the repository also welcomes non-final versions, which are then automatically linked to newer ones.

4. Your research scope is very broad; among others, you apply a digital humanist approach to the study of scientific writing in social sciences. Could you briefly describe how you conduct your research in connection with this topic?

In my postgraduate work I have been interested in how historical events are represented through mediated communication, and why only certain statements about the past are regarded as truthful representations. Currently, I've been tackling similar questions in connection with scientific writing, where I'm mostly interested in how scientists establish the validity of their claims. However, most sociological research on this topic has been purely qualitative or conducted on a handful of sampled texts. I find such an approach limited, since you can't really make general claims about whole decades of scientific writing in a particular domain based on a few dozen papers.

Consequently, I soon started wondering what a proper digital approach would reveal about this topic, and I began working on creating a corpus of Czech sociological articles from scratch. Currently, my corpus is fairly small – after the clean-up it consists of around 500 articles, but will hopefully grow with time.

5. Have there been any significant results yet?

I've obtained some interesting results by combining my corpus with a corpus of literary texts that I downloaded from the repository of Czech CLARIN. I brought the two corpora together by creating a vector space model of the documents consisting of very low-level features – the most frequent verbs that are shared between the corpora. I then applied clustering methods to the combined corpus to see which specific sociological texts have the most in common with the literary texts. As an example, clustering showed that such sociological texts often give voice to their data, by providing quotes of the people who are the subjects of the study in question. But the clusters do not only differ in language use. What I found out is that such texts are also more likely to be written by female authors, and often tend to be cited less than those texts which have little in common with fiction. Both observations turned out to be statistically significant. I plan to release this sociological corpus through the Czech CLARIN repository once it's completed.

6. Why is an infrastructure like Czech CLARIN (or CLARIN ERIC in general) important for the general research community?

I've met quite a few researchers from non-technical disciplines who oppose the use of quantitative methods in what they perceive to be qualitative research questions. I understand their point of view, which I used to share to an extent. But now that I have some experience with using language tools and resources myself, I find that such opposition often isn't really justified, although researchers must be aware of potential limitations and make sure to use the right tools for their purposes. In other words, there are many misconceptions about quantitative research and I believe that Czech CLARIN can help a lot in this regard through its user involvement events. After my personal experience of auditing the CLARIN-PLUS workshop: "Working with Digital Collections of Newspapers",[6] I think that the workshops are especially important because they're a platform where CLARIN experts can show how their tools work and how they do not only answer specific research questions from various disciplines, but also open up many approaches to doing research. An event that directly involves its participants is definitely much more convincing than a dry lecture on digital humanities that does not provide any kind of concrete examples.

Additionally, such events are often the starting points of many fruitful cross-disciplinary collaborations in which social scientists or humanities researchers team up with computer science experts. Due to such collaborations, getting involved in digital humanities does not necessarily mean that you need become an expert programmer yourself; you often only need to get intuitively acquainted with the computational methodologies and learn the basic skills, just enough to find common ground for conversations with the specialists.

7. How do your students and fellow researchers embrace the digital humanist approach? How are digital humanities in general represented in the Czech academic environment?

At the Institute of Philosophy, there is quite a lot of enthusiasm for digital humanities, since the management and many researchers see it as a step forward in scientific research. At universities, it depends a lot on the particular department. For instance, I once attended a course on programming in R that was given by Silvie Cinková from Czech CLARIN. Many students who also attended this course were from various humanities disciplines. They were very enthusiastic about learning how to programme and potentially applying programming skills to research questions within their own domains. Consequently, I think there are more students who are interested in such quantitative approaches than the management of humanities departments might realise. The problem, of course, is that the faculty at such departments doesn't usually have the required skills to teach a digital humanities course, so they often invite external teachers from the industry to teach a course or two. However, fully embracing the digital humanities would probably require a revamping of the curriculum with a greater number of digital courses tailored to topics that are directly relevant to humanities research interests.

8. What is your vision for the future of Czech CLARIN?

What I really appreciate about Czech CLARIN is that they have managed to develop tools for Czech that can easily compete with state-of-the-art language technologies developed for larger languages, like English. At LREC 2018, it was obvious to me that language technologies are rapidly becoming more and more advanced worldwide. I'm confident that Czech CLARIN will continue to keep up and make sure that their tools are always in touch with the state-of-the-art. If there's one thing that I'd like to see improved, it's the documentation of the tools and resources, which could be made more user-friendly and contain more examples of use because learning a new tool can be very intimidating.

[5] http://ufal.mff.cuni.cz/morphodita

[6] https://www.clarin.eu/event/2016/clarin-plus-workshop-working-digital-collections-newspapers

COLOPHON

This brochure is part of the 'Tour de CLARIN' volume I (publication number: CLARIN-CE-2018-1341, November 2018).

Coordinated by
Darja Fišer, Jakob Lenardič and Karolina Badzmierowska

Written by
Barbora Hladka, Darja Fišer and Jakob Lenardič

Edited by
Darja Fišer and Jakob Lenardič

Proofread by
Paul Steed

Designed by
Karolina Badzmierowska

Online version
www.clarin.eu/Tour-de-CLARIN/Publication

Publication number
CLARIN-CE-2018-1341
November 2018

Contact
CLARIN ERIC
c/o Utrecht University
Drift 10, 3512 BS Utrecht
The Netherlands
www.clarin.eu

CLARIN
Common Language Resources and
Technology Infrastructure

CLARIN 2018